Beyond 50K

Unleashing Predictive Insights into Income Classification with Machine Learning

Dominik Zabinski 306068

Agenda

- Problem definition
- Data analysis
- Data preparation
- Methods
- Results
- Summary

Problem definition

- Which features are most important while predicting salary?
- What is the best model for this task?
- How good is the best model?

Data analysis

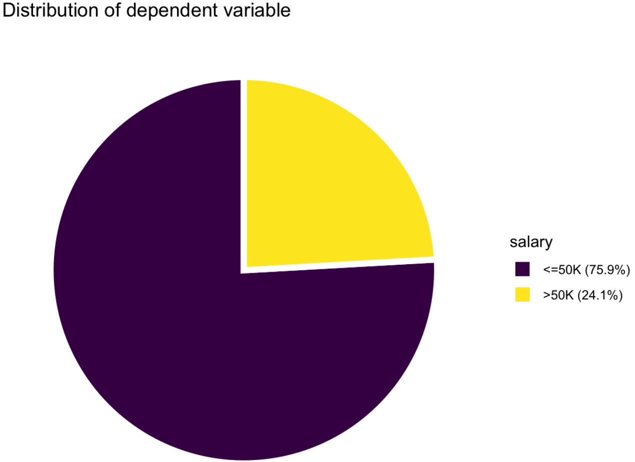

Figure 1: Salary distribution

Variable	Min	Avg	Sd	Max	# miss.	# dist.
age	17.00	38.58	13.63	90.00	0	73
capital.gain	0.00	1070.19	7351.82	99999.00	0	119
capital.loss	0.00	85.56	398.37	4356.00	0	92
education.num	1.00	10.09	2.57	16.00	0	16
feat01	0.15	1.03	0.31	1.92	0	42561
feat02	0.00	0.53	0.12	1.00	0	42561
feat03	0.14	1.04	0.31	1.95	0	42561
feat04	0.00	0.53	0.11	1.00	0	42561
feat05	0.09	1.01	0.32	1.94	0	42561
feat06	0.00	0.55	0.12	1.00	0	42561
feat07	0.14	0.98	0.31	1.87	0	42561
feat08	0.10	1.02	0.31	1.91	0	42561
feat09	0.13	0.98	0.31	1.82	0	42561
feat10	0.04	1.02	0.32	1.92	0	42561
fnlwgt	12285.00	189412.02	105635.52	1484705.00	0	21648
hours.per.week	1.00	40.44	12.33	99.00	0	94

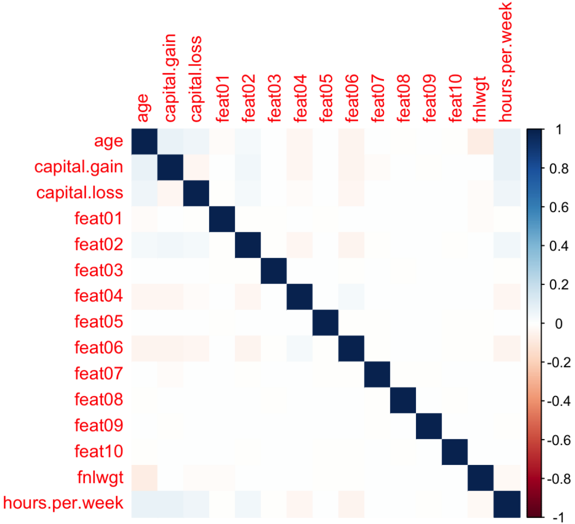

Figure 2: Correlation matrix between continous variables

Variable	# dist.	# miss.	Top levels	Dist. range	# range
education	16	0	HS-grad (32.18); Some-college (22.41); Bachelors (16.62)	0.14% - 32.18%	61 - 13697
marital.status	7	0	Married-civ-spouse (45.92); Never-married (32.69); Divorced (13.74)	0.07% - 45.92%	29 - 19542
native.country	42	0	United-States (89.54); Mexico (1.96); ? (1.76)	0% - 89.54%	1 - 38111
occupation	15	0	Prof-specialty (12.8); Exec-managerial (12.47); Craft-repair (12.43)	0.02% - 12.8%	10 - 5448
race	5	0	White (85.37); Black (9.63); Asian-Pac-Islander (3.23)	0.82% - 85.37%	351 - 36335
relationship	6	0	Husband (40.41); Not-in-family (25.5); Own-child (15.53)	2.97% - 40.41%	1266 - 17200
sex	2	0	Male (66.86); Female (33.14)	33.14% - 66.86%	14106 - 28455
workclass	9	0	Private (69.73); Self-emp-not-inc (7.75); Local-gov (6.47)	0.03% - 69.73%	11 - 29679

Data preparation

- joining small groups in discrete variable (data-driven and/or gut feeling)
- one-hot encoding discrete variables
- split training/testing in 70/30 proportion
- normalize continuous variables
- 5-fold cross-validation as sampling method

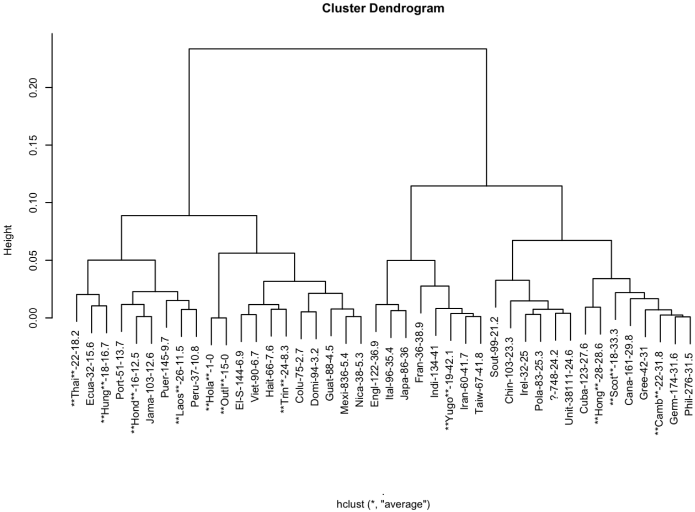

Figure 3: Dendrogram for ‘native.country’ variable

Methods

Single models:

- Decision tree
- Random forest
- Generalized Boosted Regression Modeling (GBM)
- eXtreme Gradient Boosting (XGBoost)
- Logistic regression (as benchmark)

Ensemble model: weighted combination of single models

Stacked models:

- Logistic Regression as top layer model
- XGBoost as top layer model

Assessment metric: AUC

Results

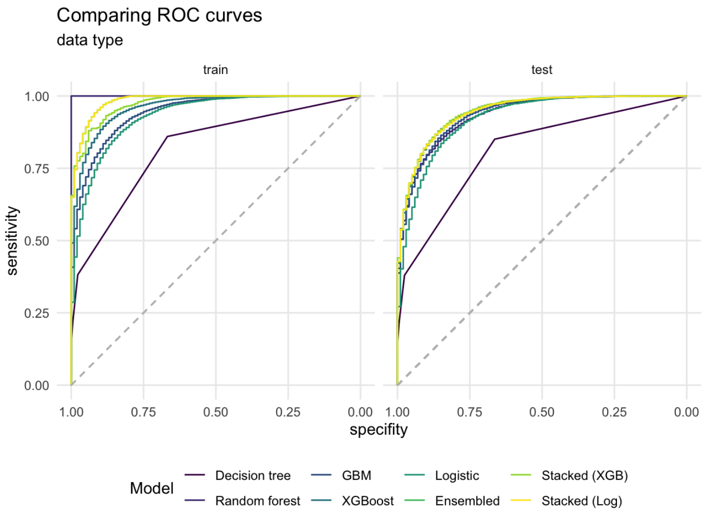

Figure 4: Comparison of ROC curves - data type

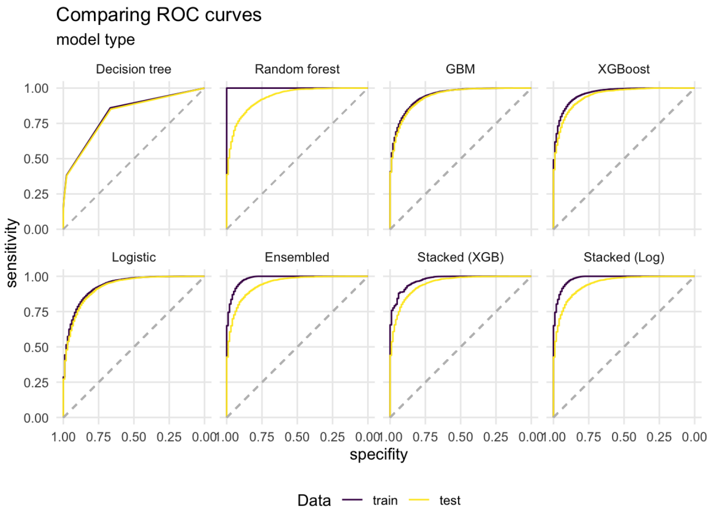

Figure 5: Comparison of ROC curves - model type

Model	AUC (Train)	AUC (Test)	Gini (Train)	Gini (Test)
GBM	94.34%	93.6946%	88.7%	87.389%
Decision tree	81.93%	81.2647%	63.9%	62.529%
Logistic	93.09%	92.2270%	86.2%	84.454%
XGBoost	96.13%	94.4457%	92.3%	88.891%
Stacked (XGB)	97.35%	94.6237%	94.7%	89.247%
Ensembled	98.35%	94.4524%	96.7%	88.905%
Stacked (Log)	98.35%	94.4524%	96.7%	88.905%
Random forest	**100.00%**	**93.3522%**	**100.0%**	**86.704%**

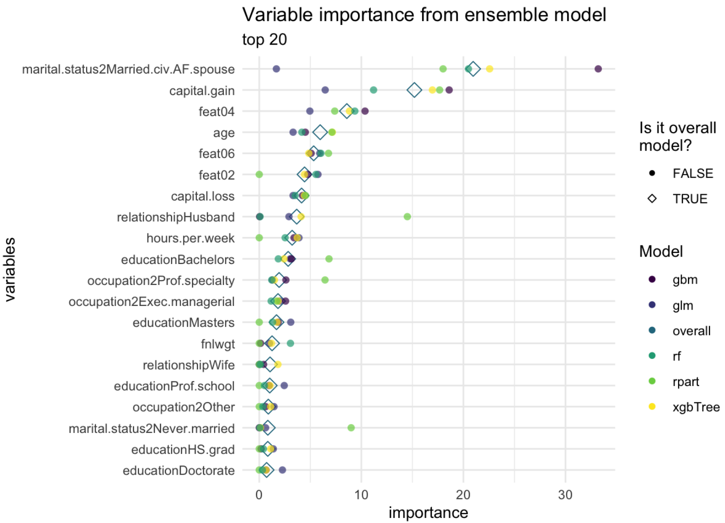

Figure 6: Variable importance - ensemble model

Summary

- Which features are most important while predicting salary?
  - being married and living with a spouse
  - capital gain
  - age
  - feat02, feat04 and feat06
  - higher education (BA+)
- What is the best model for this task?
  - Stacked with XGBoost as top layer model,
  - **on production** I'd consider single XGBoost
- How good is the best model?
  - AUC: 94%
- What could be done differently?
  - hyperparameter tuning
  - binning continuous variables